SCORE: Soft Label Compression-Centric Dataset Condensation via Coding Rate Optimization

Bowen Yuan* Yuxia Fu * Zijian Wang Yadan Luo Zi Huang
The University of Queensland
{bowen.yuan, yuxia.fu, zijian.wang, y.luo, helen.huang}@uq.edu.au

A. Technical Appendices and Supplementary Material

This supplementary material provides additional descriptions of proposed dataset condensation method SCORE, including extensional experiments, theoretical proofs and empirical details. Visualizations of condensed datasets are demonstrated to enhance understanding of the proposed method.

- **Section A.1.1:** Edge Case Study.
- **Section A.1.2:** Training Effect of Soft Label Compression.
- **Section A.1.3:** Storage Budget of Soft Labels.
- **Section A.2:** Limitation and Future Work.
- **Section A.3.1:** Proof of Coding Rate for Rank Approximation.
- **Section A.3.2:** Proof of Coding Rate Submodularity.
- **Section A.4:** Unified Criterion Derivation.
- **Section A.5:** Hyper-Parameter Settings.
- **Section A.6:** Visualizations.

A.1. Extensional Experiments

A.1.1. Edge Case Study

We evaluate SCORE under extreme soft label compression scenarios. We test SCORE with a large compression ratio of 40× on ImageNet-1K using both 10 and 50 IPC, as shown in Table 1. SCORE consistently achieves superior accuracy, attaining 31.2% at 10 IPC and 52.6% at 50 IPC, while maintaining storage overhead of 0.3GB and 1.4GB respectively.

In addition, we investigate performance at extremely low and high IPC settings. We condense ImageWoof and ImageNette to both 1 IPC and 100 IPC settings. As shown in Table 2, SCORE demonstrates better performance across the edge scenarios. With minimal data (1 IPC), SCORE achieves 17.9% accuracy on ImageWoof and 29.5% on ImageNette. While the performance on ImageNette notably increased at 100 IPC, reaching 86.6%, the performance on ImageWoof remained relatively stable, with SCORE achieving 75.1%. The results highlight effectiveness of SCORE across diverse datasets and compression ratios, even in the edge cases.

Table 1. Performance comparison under extreme compression ratio (40×) on ImageNet-1K.

IPC	SCORE (Ours)	LPLD [3]	Storage
10	**31.2 ± 0.1**	20.2 ± 0.3	0.3G
50	**52.6 ± 0.3**	46.7 ± 0.3	1.4G

A.1.2. Training Effect of Soft Label Compression

To investigate the effect of soft label compression on the model training process, we conducted extensive experiments using ImageNet-1K with 5 IPC. We compared standard model training with uncompressed soft labels against training with soft label

*The authors contribute equally to the research.

Table 2. Performance comparison under extreme IPC settings

Dataset	IPC	SCORE (Ours)	RDED [2]
ImageWoof	1	$\mathbf{17.9 \pm 0.8}$	17.9 ± 1.0
	100	$\mathbf{75.1 \pm 0.2}$	75.1 ± 0.2
ImageNette	1	$\mathbf{29.5 \pm 0.7}$	27.7 ± 1.1
	100	$\mathbf{86.6 \pm 0.5}$	81.3 ± 1.3

compression. As shown in Figure 1, while soft label compression introduces some information loss with higher loss values and marginally lower accuracy, both approaches follow similar learning trajectories and convergence patterns. Despite the moderate performance trade-off, soft label compression preserves the essential learning signals within a limited storage budget.

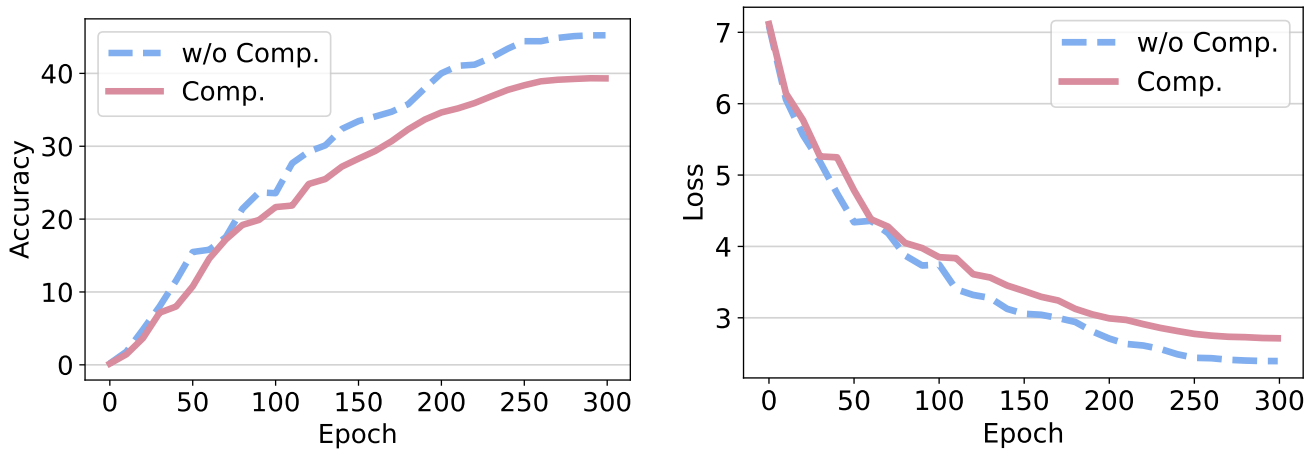

Figure 1. Comparison of model performance with and without compression across training epochs on ImageNet-1K. **Left**: accuracy comparison. **Right**: loss comparison.

A.1.3. Storage Budget of Soft Labels After Compression

In this section, we analyze the storage budget of soft labels under different compression ratios on both ImageNet-1K and Tiny-ImageNet, as summarized in Table 3. The compression ratio of $1\times$ refers to the uncompressed soft labels generated after 300 epochs of training. As shown in the table, the storage requirement can be substantial under large-scale settings. For example, soft labels for ImageNet-1K at 50 IPC occupy as much as 56 GB. This contradicts the core objective of dataset distillation, which aims to reduce storage and computation while preserving model performance. By applying our soft label compression method with ratios of $10\times$, $20\times$, and $30\times$, the storage cost is reduced significantly to 5.6 GB, 2.8 GB, and 1.9 GB, respectively. This demonstrates that our approach substantially reduces the storage budget for soft labels, making it more practical for large-scale distillation scenarios while still preserving the effectiveness of the synthetic dataset.

A.2. Limitation and Future Work

Currently, we employ a greedy algorithm for dataset condensation. Even though SCORE achieves competitive results in extremely low IPC settings such as 1 IPC, this greedy approach can still lead to sub-optimal solutions, as individual sample selection has an important impact on the overall performance. In future work, we aim to investigate more effective approaches to address low IPC challenges. Additionally, extending dataset condensation beyond classification to other applications represents a promising direction for future exploration.

Table 3. Storage budget of soft labels for ImageNet-1K and Tiny-ImageNet at 10 and 50 IPC under various compression ratios.

Dataset	IPC	Compression Ratio			
		1×	10×	20×	30×
Tiny-ImageNet	10	460.0M	46.0M	23.0M	15.3M
	50	2.3G	236.0M	118.0M	78.7M
ImageNet-1K	10	11.6G	1.2G	0.6G	0.3G
	50	56.0G	5.6G	2.8G	1.9G

A.3. Proof

A.3.1. Proof of Lemma 1

Log-det is justified as a smooth approximation for the rank function [1]. For any matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$, the coding rate function is strictly concave w.r.t. $\mathbf{Z}\mathbf{Z}^\top$ and provides a smooth approximation to the rank function.

Proof. For all vectors $\mathbf{x}$, given a coefficient λ and feature matrix $\mathbf{Z}$, we have

$$\begin{aligned}\mathbf{x}^\top(\mathbf{I}+\lambda\mathbf{Z}\mathbf{Z}^\top)\mathbf{x} &= \mathbf{x}^\top\mathbf{I}\mathbf{x}+\lambda\mathbf{x}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{x}\\ &= \mathbf{x}^\top\mathbf{x}+\lambda\mathbf{x}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{x}\\ &= \|\mathbf{x}\|^2+\lambda\|\mathbf{Z}^\top\mathbf{x}\|^2\\ &\geq 0.\end{aligned} \tag{1}$$

Therefore, the term $\mathbf{I}+\lambda\mathbf{Z}\mathbf{Z}^\top$ is positive semidefinite. Now let $\mathbf{Z}\mathbf{Z}^\top=\mathbf{M}$. Assume $\mathbf{M}_1$ and $\mathbf{M}_2$ are positive semidefinite matrices, and let $0\leq\alpha\leq 1$. Define:

$$\begin{aligned}X_1 &= \mathbf{I}+\lambda\mathbf{M}_1,\\ X_2 &= \mathbf{I}+\lambda\mathbf{M}_2.\end{aligned}$$

For any convex combination:

$$\alpha X_1+(1-\alpha)X_2 = \alpha(\mathbf{I}+\lambda\mathbf{M}_1)+(1-\alpha)(\mathbf{I}+\lambda\mathbf{M}_2) \tag{2}$$

$$= \mathbf{I}+\lambda(\alpha\mathbf{M}_1+(1-\alpha)\mathbf{M}_2). \tag{3}$$

$\log\det(X)$ is concave on the set of positive definite matrices. Therefore:

$$\log\det(\alpha X_1+(1-\alpha)X_2)\geq\alpha\log\det(X_1)+(1-\alpha)\log\det(X_2) \tag{4}$$

Substituting to the definitions:

$$\log\det(\mathbf{I}+\lambda(\alpha\mathbf{M}_1+(1-\alpha)\mathbf{M}_2))\geq\alpha\log\det(\mathbf{I}+\lambda\mathbf{M}_1)+(1-\alpha)\log\det(\mathbf{I}+\lambda\mathbf{M}_2). \tag{5}$$

This confirms that $f(\mathbf{M})=\log\det(\mathbf{I}+\lambda\mathbf{M})$ is concave in $\mathbf{M}$.
Let $\mathbf{M}$ have eigenvalues $\mu_1,\mu_2,\ldots,\mu_n$ (all non-negative since $\mathbf{M}$ is positive semidefinite).
Then $\mathbf{I}+\lambda\mathbf{M}$ has eigenvalues $1+\lambda\mu_1,1+\lambda\mu_2,\ldots,1+\lambda\mu_n$.
The determinant is the product of eigenvalues, so:

$$\det(\mathbf{I}+\lambda\mathbf{M})=\prod_{i=1}^{n}(1+\lambda\mu_i) \tag{6}$$

Taking the logarithm:

$$\log\det(\mathbf{I}+\lambda\mathbf{M})=\sum_{i=1}^{n}\log(1+\lambda\mu_i) \tag{7}$$

Now, as $\lambda\to\infty$:

- For $\mu_i > 0$: $\log(1 + \lambda\mu_i) \approx \log(\lambda\mu_i) = \log(\lambda) + \log(\mu_i)$
- For $\mu_i = 0$: $\log(1 + \lambda\mu_i) = \log(1) = 0$

Let $r = rank(\mathbf{M})$, which is the number of positive eigenvalues. For a non-zero λ, we have:

$$\log\det(\mathbf{I} + \lambda\mathbf{M}) \approx \sum_{i:\mu_i>0} \log(\lambda\mu_i) \tag{8}$$

$$= \sum_{i:\mu_i>0} \log(\lambda) + \sum_{i:\mu_i>0} \log(\mu_i) \tag{9}$$

$$= r\log(\lambda) + \sum_{i:\mu_i>0} \log(\mu_i) \tag{10}$$

Thus, $\log\det(\mathbf{I} + \lambda\mathbf{M}) \approx r\log(\lambda) + c$, where $c = \sum_{i:\mu_i>0} \log(\mu_i)$ is a constant that depends on the non-zero eigenvalues of $\mathbf{M}$.

This shows that for large λ, minimizing $\log\det(\mathbf{I} + \lambda\mathbf{M})$ is approximately equivalent to minimizing the rank of $\mathbf{M}$, up to scaling by $\log(\lambda)$ and an additive constant.

□

A.3.2. Proof of Lemma 2

Given a set of features $\mathbf{Z}$, coding rate function $R_{\mathrm{I}}(\mathbf{Z}) = \log\det(\mathbf{I} + \lambda\mathbf{Z}\mathbf{Z}^\top)$ satisfies the definition of submodular function. To prove submodularity, R_{IC} should satisfy diminishing returns property: for any sets $\mathcal{A} \subseteq \mathcal{B}$ and any element $i \notin \mathcal{B}$, if a function f is submodular, we have $f(\mathcal{A} \cup \{i\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{i\}) - f(\mathcal{B})$. Let $\mathcal{A} \subseteq \mathcal{B}$ be two sets of indices for $\mathbf{Z}$, we need to show:

$$R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{A}\cup\{i\}}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{A}}) \geq R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{B}\cup\{i\}}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{B}}). \tag{11}$$

Proof. Given a set $\mathcal{S}$, we can rewrite the term using matrix determinant lemma:

$$\begin{aligned} R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{S}} \cup \{i\}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{S}}) &= \log\det(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{S}\cup\{i\}}\mathbf{Z}_{\mathcal{S}\cup\{i\}}^\top) - \log\det(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^\top) \\ &= \log\frac{\det(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{S}\cup\{i\}}\mathbf{Z}_{\mathcal{S}\cup\{i\}}^\top)}{\det(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^\top)} \\ &= \log(1 + \lambda\mathbf{z}_i^\top(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^\top)^{-1}\mathbf{z}_i). \end{aligned} \tag{12}$$

Since $\mathcal{A} \subseteq \mathcal{B}$, thus $\mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top$ is a principal submatrix of $\mathbf{Z}_{\mathcal{B}}\mathbf{Z}_{\mathcal{B}}^\top$, thus we have:

$$\mathbf{Z}_{\mathcal{B}}\mathbf{Z}_{\mathcal{B}}^\top - \mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top \succeq 0, \tag{13}$$

where $\succeq$ denotes the Löwner partial order. Using matrix inversion lemma, we can show that:

$$\begin{aligned} (\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top)^{-1} - (\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{B}}\mathbf{Z}_{\mathcal{B}}^\top)^{-1} &\succeq 0 \\ \mathbf{z}_i^\top[(\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top)^{-1} - (\mathbf{I} + \lambda\mathbf{Z}_{\mathcal{B}}\mathbf{Z}_{\mathcal{B}}^\top)^{-1}]\mathbf{z}_i &\geq 0. \end{aligned}$$

Therefore, $R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{J}} \cup \{i\}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{J}})$ is monotonically decreasing, which proves that: given $\mathcal{A} \subseteq \mathcal{B}$, the inequality holds:

$$R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{A}} \cup \{i\}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{A}}) \geq R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{B}} \cup \{i\}) - R_{\mathrm{I}}(\mathbf{Z}_{\mathcal{B}}), \tag{14}$$

and R_{I} is a submodular function. □

A.4. Derivation for Unified Criterion

In the image selection process, we aim to select samples that maximize the marginal gain of the unified criterion in each round. For a selected set $\mathcal{S}$, when a function f_{sub} is submodular, the marginal gain of adding element i to $\mathcal{S}$ is defined as:

$$f_{\mathrm{sub}}(\{i\} \mid \mathcal{S}) = f_{\mathrm{sub}}(\mathcal{S} \cup \{i\}) - f_{\mathrm{sub}}(\mathcal{S}). \tag{15}$$

Based on this submodularity property, given that certain terms remain constant during a single selection iteration, the optimal image $\mathbf{x}^*$ can be determined as follows:

$$\begin{aligned} x^* &= \arg\max_x R(\mathbf{x}, \mathbf{Y} \mid \mathcal{X}') \\ &= \arg\max_x R_{\mathrm{I}}(f(\{\mathbf{x}\} \cup \mathcal{X}'; \theta_h)) - R_{\mathrm{I}}(f(\mathcal{X}'; \theta_h)) \\ &\quad - \alpha R_{\mathrm{D}}(f(\{\mathbf{x}\} \cup \mathcal{X}'; \theta_h)) + \alpha R_{\mathrm{D}}(f(\mathcal{X}'; \theta_h)) - \beta R_{\mathrm{C}}(\mathbf{Y}) \\ &= \arg\max_x R_{\mathrm{I}}(f(\{\mathbf{x}\} \cup \mathcal{X}'; \theta_h)) - \alpha R_{\mathrm{D}}(f(\{\mathbf{x}\} \cup \mathcal{X}'; \theta_h)) - \beta R_{\mathrm{C}}(\mathbf{Y}). \end{aligned} \tag{16}$$

A.5. Hyper-Parameter Settings

All experiments utilize the AdamW optimizer for model training and apply *RandomResizeCrop*, *RandomHorizontalFlip*, and *CutMix* as data augmentations. The specific hyperparameter settings for all datasets are detailed in Table 4. These include the values of α and β used for image selection, the minimum and maximum scales for the *RandomResizeCrop* augmentation, as well as other network parameters such as learning rate, weight decay, temperature T, batch size, and the number of training epochs.

Table 4. Hyper-parameters for ImageNet-1K, Tiny-ImageNet, ImageWoof and ImageNette.

	ImageNet-1K		Tiny-ImageNet		ImageWoof		ImageNette	
	IPC		IPC		IPC		IPC	
Hyper-parameter	10	50	10	50	10	50	10	50
α	5	5	5	10	5	5	0.1	0.1
β	1	1	0.001	0.001	0.1	0.1	0.1	0.1
Min scale of Aug	0.4	0.4	0.3	0.4	0.3	0.3	0.2	0.2
Max scale of Aug	1	1	1	1	1	1	1	1
Learning Rate	0.001	0.001	0.001	0.001	0.001	0.001	0.0005	0.0005
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
T	20	20	20	20	20	20	10	10
Batch Size	64	64	50	50	64	64	10	10

A.6. Visualization

The visualizations of our selected images, presented in Figures 2 to 5, are randomly sampled from the condensed datasets of ImageNet-1K, Tiny-ImageNet, ImageWoof, and ImageNette, respectively.

Figure 2. Visualization of the condensed ImageNet-1K.

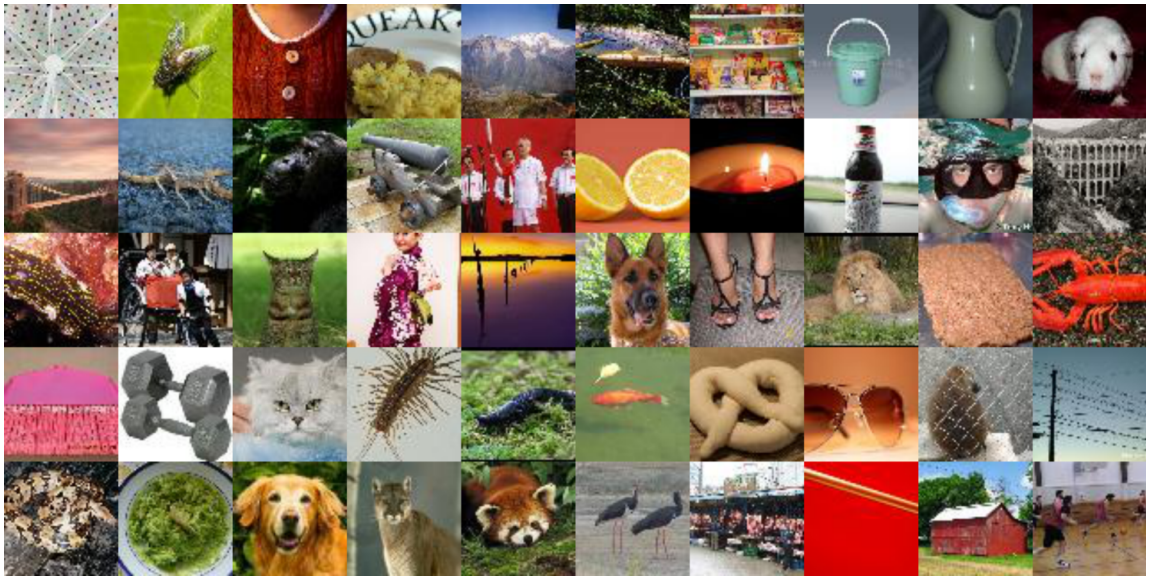

Figure 3. Visualization of the condensed Tiny-ImageNet.

Figure 4. Visualization of the condensed ImageWoof.

Figure 5. Visualization of the condensed ImageNette.

References

[1] Maryam Fazel, Haitham A. Hindi, and Stephen P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *American Control Conference, ACC 2003, Denver, CO, USA, June 4-6 2003*. IEEE, 2003. 3

[2] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024. 2

[3] Lingao Xiao and Yang He. Are large-scale soft labels necessary for large-scale dataset distillation? In *NeurIPS*, 2024. 1